21｜持续精进：如何在机器学习领域中找准前进的方向？

2021-10-15 黄佳

《零基础实战机器学习》 课程介绍 >

讲述：黄佳
时长 24:22　大小 22.32M

你好，我是黄佳。欢迎来到零基础实战机器学习。

我常常说，一旦你踏进了机器学习领域，就等同于踏进了“终身学习”之旅。因为机器学习领域庞大而繁杂，我们这 20 多讲的内容仅仅是一个开始。除了我们讲过的监督学习、无监督学习、半监督学习、强化学习、深度学习和集成学习之外，还有迁移学习、自动化机器学习、联邦学习、图神经网络等多种学习类型和研究方向。

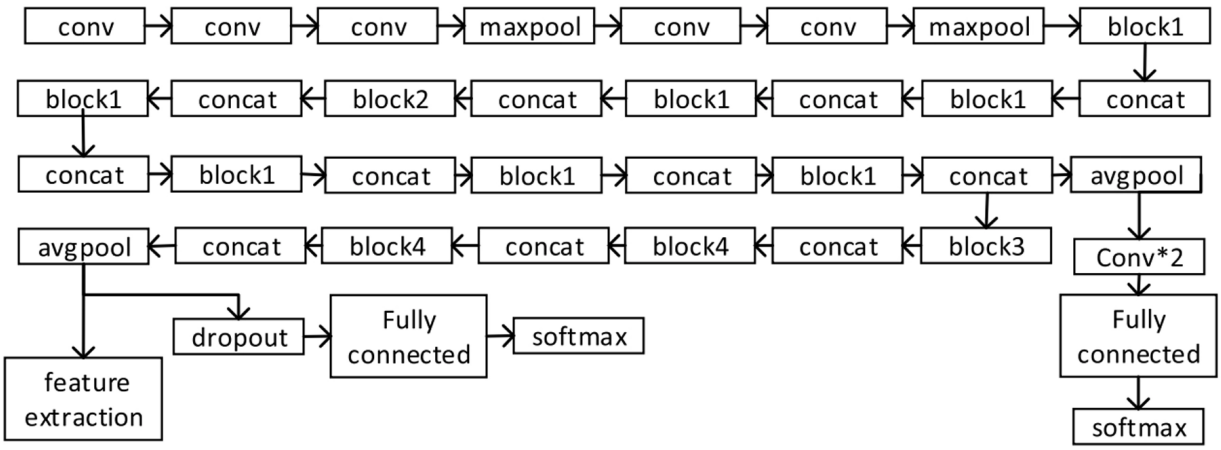
那么，作为机器学习使用者的我们，面对机器学习领域里众多的“新热点”和种类繁多的研究方向，应该怎么去找准关注点和发力点，更高效地赋能自己的个人发展呢？在这一中，我将结合机器学习最新发展现状和技术演进趋势，来带你探讨这个问题。

迁移学习

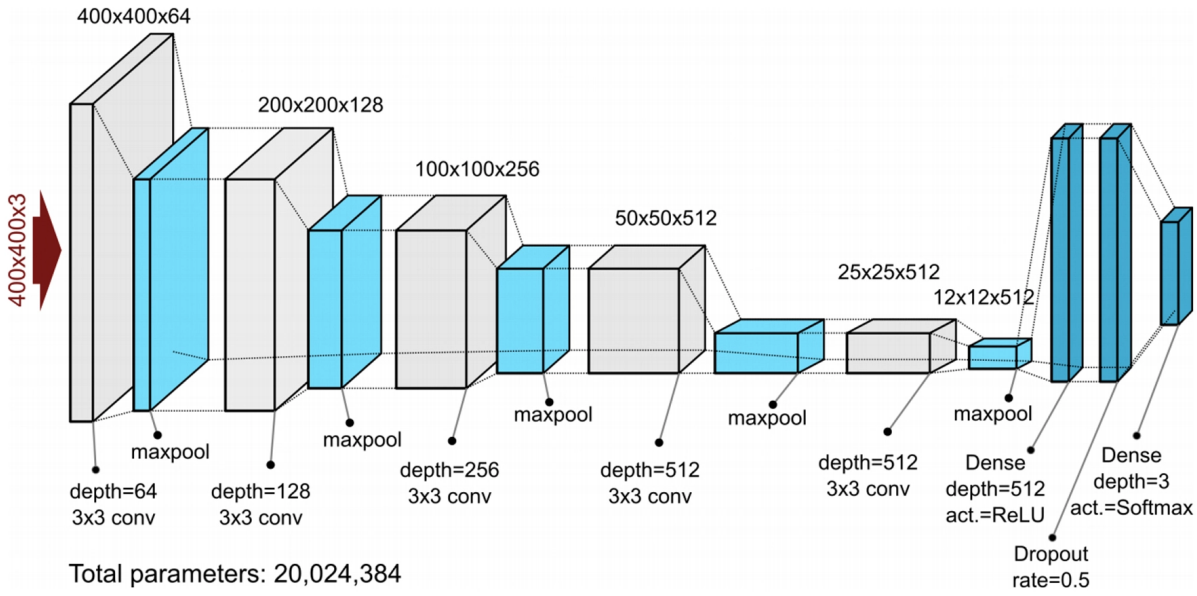
我这里给你的第一个建议就是**不要重复造轮子**。要知道，人类的知识大厦都是经年累月，逐层搭建起来的，机器学习发展到今天，不仅有非常多的包供我们调用，甚至也已经有越来越多的模型，我们可以直接使用。在上一讲中，我们就预训练了一个模型，然后在 Web 应用中直接把这个训练好的模型拿过来使用。其实，对于别人已经预训练成功的大型模型，我们也可以做个微调，然后应用到自己的业务场景中。这个思路，就是“迁移学习”。

迁移学习（Transfer Learning）是把为任务 A 开发的模型作为初始点，重新使用在为任务 B 开发模型的过程中。在深度学习的计算机视觉任务和自然语言处理任务里，迁移学习特别常见。因为大型神经网络模型的训练、调试、优化的过程会消耗巨大的时间资源和计算资源，为了避免资源浪费，很多公司会找到开源发布出来的模型，通过迁移学习将这些模型已习得的强大技能迁移到相关的的问题上。

举例来说，我们在第 11 讲中，创建了一个简单的 CNN 神经网络的鲜花分类器，准确率达到了 80% 左右。而这种深度学习图像分类神经网络，许多知名机构其实也在不断地发布，它们更复杂，精度也更高，比如牛津 VGG 模型、谷歌 Inception 模型、微软 ResNet 模型等。下面两张图，展示的就是 Google Inception V3 模型和 VGG19 模型的网络结构。



Google Inception V3 结构（来源：Pulmonary Image Classification Based on Inception-v3 Transfer Learning Model ）



牛津VGG 19的网络结构（来源：Melanoma Thickness Prediction Based on Convolutional Neural Network with VGG-19 Model Transfer Learning）

作为机器学习的应用者，我们就可以利用这些训练好的大型卷积网络，来实现自己的机器学习任务，而且这个迁移过程并没有我们想象得那么困难。我们现在就试着对牛津 VGG19 的网络模型基础进行微调。

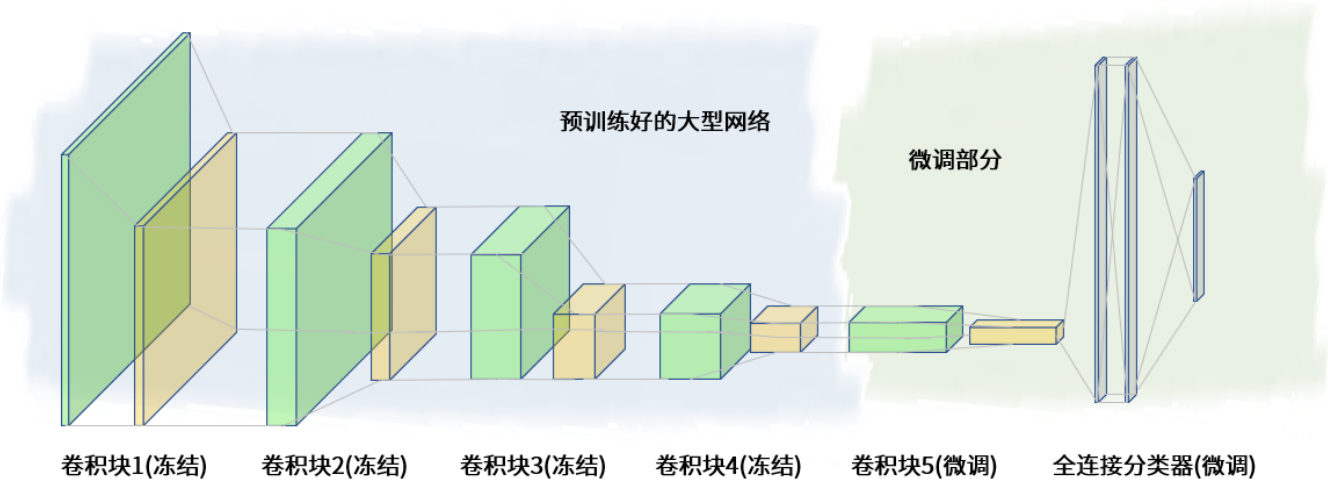
复制代码

```
import tensorflow as tf # 导入tensorflow
from tf.keras.applications import VGG19 # 基网络是VGG19
from tf.keras import models # 导入模块
from tf.keras import layers # 导入层
from tf.keras import optimizers # 导入优化器
# 预训练的卷积基
conv_base = VGG19(weights='imagenet', include_top=False,
input_shape=(150, 150, 3))
conv_base.trainable = True # 解冻卷积基
# 冻结其他卷积层，仅设置block5_conv1 可训练
set_trainable = False
for layer in conv_base.layers:
    if layer.name == 'block5_conv1': #仅设置block5_conv1 可微调
        set_trainable = True
    if set_trainable:
        layer.trainable = True #仅设置block5_conv1 可微调
    else:
        layer.trainable = False #其它4个块冻结
model = models.Sequential() # 序贯模型
model.add(conv_base) # 基网络的迁移
model.add(layers.Flatten()) # 展平层
model.add(layers.Dense(128, activation='relu')) # 微调全连接层
model.add(layers.Dense(10, activation='softmax')) # 微调分类输出层
model.compile(loss='binary_crossentropy', # 交叉熵损失函数
```

```
# 为优化器设置小的学习速率，就是在微调第5 卷积层的权重
optimizer=optimizers.adam(lr=1e-4),
metrics=['acc']) # 评估指标为准确率
model.fit(X_train, y_train, epochs=2, validation_split=0.2) # 训练网络
```

上面的代码并不难理解，其中最主要的逻辑就是，我们要先解冻 VGG19 所包含的 5 个卷积块的基模型，然后再通过 set_trainable 对象冻结其中 4 个卷积块，这就是我们从 VGG19 这个模型中迁移过来的知识，最后我们把 VGG19 的第 5 个模块 block5_conv1 设为可调、可训练的模块即可。

这样，我们就实现了在预训练的卷积网络 VGG19 基础之上的微调。微调之后，我们得到了一个新模型：



基于VGG19 网络的迁移学习模型

有了这个新模型后，我们就可以训练新模型，并拿这个新模型对任意图像数据集做分类了，当然也包括我们的易速鲜花图片数据。

通过上面的迁移学习示例，我希望你能具有“多利用他人成果，避免从零开始”的思维，当大型的模型有所发展，我们就可以在新模型的基础上“与时俱进”，把最新的前沿成果利用起来。

接下来，我要给你第二个建议，**多关注“自动化机器学习”的可能性**。

自动化机器学习

你肯定已经体会到了特征工程的麻烦和参数调优的苦恼。在典型的机器学习应用中，为了让数据适合机器学习，机器学习工程师必须应用适当的数据预处理、特征工程、特征提取和特征选择方法。在这些步骤之后，还必须进行算法选择和超参数优化，以最大限度地提高模型的预测性能。

这里面的每一步都非常有挑战。不仅如此，在尝试部署新开发的 AI 系统和机器学习模型时，企业和组织也经常在系统可维护性、可扩展性和治理方面苦苦挣扎，而 AI 项目往往又无法产生预期的回报。

那如何让 AI 项目更加顺畅地落地呢？自动化机器学习（AutoML）非常值得一提。自动化机器学习是把机器学习应用于现实世界问题，并将任务自动化的过程。它力图简化机器学习项目中的步骤，形成从数据预处理、特征工程，到算法选择，参数调优直至部署上线的完整解决方案，使整个机器学习项目变得更加高效。AutoML 的目标就是让非专家也能使用机器学习模型和技术。

当调参和部署的负担减轻之后，机器学习工程师们就可以花更多的时间来关注机器学习模型本身，尝试不同的架构，设计并优化新的算法。

自动化机器学习包括以下三个方面。这里我在介绍每一种自动化机器学习的同时，也给出了相应的工具的链接，如果你有兴趣，可以尝试尝试。

1. **自动化特征工程**

我们知道，机器学习算法的性能很大程度上取决于数据特征的质量，所以特征工程涉及了大量的试验和深厚的领域知识，对于新手来说，要进行有效的特征工程，难度就比较大。而自动化特征工程的目的就是通过不断迭代、调整，自动创建出新的特征集，直到模型达到令人满意的准确性为止。

目前已存在的一些自动化特征工程框架包括 DataRobot、H2O 的 Driverless AI和 tsfresh 库，这些框架可以辅助我们探索特征之间的关系，完成一部分特征工程的任务。

2. **自动化模型选择和超参数调整**

机器学习算法的目的是根据数据特征来训练机器模型，以预测新数据的标签值。但每种算法都适合于特定类型的问题，没有一种机器学习算法在所有数据集上都能有最佳表现，而且超参数调整过程也必不可少。在模型选择过程中，我们会尝试使用不同的变量，不同的系数或不同的超参数。这个过程极大地考验着工程师的经验、技巧和耐心。

这时候，我们可以就借助自动化模型选择工具，针对手头的特定任务，遍历所有合适的模型，并选择产生最高准确性或最低误差的模型。目前市面上已有的自动化模型选择工具有Auto-sklearn、H2O 的Driverless AI、微软 Azure 的Automated ML工具包、Google Cloud 的AutoML等，这些工具只要有了贴好标签的数据后，就可以直接训练模型，帮我们自动尝试不同的模型，并完成参数调试和算法优化。

3. **自动部署机器学习模型**

虽然机器学习工程师会把项目焦点放在开发算法和调试参数上，模型的部署和在生产系统中的运行维护很少被提及，但是，这里面也存在很多挑战。有时候，部署上线花费的时间甚至会超过训练、调试模型所用的时间。因此，自动部署的工具，对于并不了解 DevOps 的机器学习工程师们来说，其实大有用途。

我们常用的自动部署工具有：

- Microsoft Machine Learning Service：可以将机器学习模型部署为可扩展的 Kubernetes 集群的 Web 服务；
- Amazon Sage Maker和Google Cloud ML：能快捷实现 Web 部署功能。通过对托管机器学习模型的 Web 服务进行 HTTP 调用，这二者都可以实现推断和预测数据；
- TensorFlow Serving：可将 Tensorflow 机器学习模型部署到生产系统中。只需要几行代码，我们就可以通过 Tensorflow 模型生成 API。

从上述三个方面可以看到，从特征工程到模型选择、参数调试，再到生产部署，自动化机器学习几乎覆盖了机器学习流程中的每一个环节。当然，这些工具是不是好用，能不能真正解决机器学习项目中的痛点，就要仁者见仁智者见智了。也许这些自动化工具还处于它们生命周期的初级阶段，不是我们所希望的那么“易学易用”。但是，我们有理由认为，优秀的 AutoML 工具一定会让机器学习变得越来越简单易用。

接下来，我要给你的第三个，也是最后一个建议：**关注并思考机器学习的最新进展将如何进一步拓展 AI 落地应用的覆盖面。**因为这样能够帮你占得先机，一旦实用的工具包出现，技术开始实际落地时，你就可以先人一步把它们用起来，做潮流的引领者。而目前，最新机器学习热门研究方向是图神经网络。

图神经网络

图神经网络（GNN）是 2021 年机器学习学术界的“红星”，也是各大深度学习顶会研究的焦点技术，到现在依旧方兴未艾。

我们说，迁移学习解决的主要问题是如何在机器学习模型之间进行知识的传递和迁移，而自动化机器学习的愿景是为了让机器学习的应用落地过程变得越来越简单，那么，深度强化学习和图神经网络等学习类型的出现，则意味着**人工智能的应用领域变得越来越广阔，而且多种 AI 源流正在融合，让机器学习领域呈现出百川归海的趋势。**

从机器学习的源流演变历程来看，有三个主要学派起着关键作用，它们分别是符号主义（Symbolicism）、连接主义（Connectionism）和行为主义（Actionism）。

符号主义注重研究知识表达和逻辑推理。这一学派的主要成果有两个，一个是贝叶斯因果网络，另一个是知识图谱。贝叶斯因果网络又称信度网络，是 Bayes 方法的扩展，也是目前知识表达和推理领域最有效的理论模型之一。它的提出者是 2011 年图灵奖获得者 Judea Pearl。而知识图谱主要由谷歌、微软、百度等搜索引擎公司推动，目标是把搜索引擎从关键词匹配，推进到语义匹配。

连接主义的起源是仿生学，用数学模型来模仿神经元，形成深层神经网络模型。神经网络在 1950 年代诞生，期间也有过发展的低迷期，但目前深度学习风头正劲，领军人物有 Yann LeCun、Geoff Hinton 和 Yoshua Bengio，人称三巨头，他们在 2018 年同获图灵奖。不过，我们说过，深度学习模型最遭人诟病的缺陷是解释性差。Judea Pearl 就曾经在《为什么》一书中诟病深度学习既不易解释，又无法进行因果推理，你如果感兴趣可以了解一下。

而**行为主义**则是把控制论引入到机器学习中，这一学派认为，机器的自适应、自组织、自学习功能是由系统的输入输出反馈行为决定的。其中，最著名的成果是强化学习。强化学习，强调如何基于环境而行动，以取得最大化的预期利益，其灵感来源于心理学中的行为

主义理论，即有机体如何在环境给予的奖励或惩罚的刺激下，逐步形成对刺激的预期，产生能获得最大利益的习惯性行为。

从几年前开始，Google 的 DeepMind 研究员们把传统强化学习与深度学习融合，形成了**深度强化学习**（deep reinforcement learning）。2013 年，DeepMind 发布的论文《用深度强化学习打 Atari 游戏》（Playing Atari with Deep Reinforcement Learning）就是深度强化学习的开山之作，它介绍深度强化网络 DQN（deep Q network）集成了强化学习的决策能力，也集成了深度学习在复杂问题问题上强大的拟合能力。

深度强化学习的出现让强化学习技术真正走向实用。到目前为止，深度强化学习领域出现了大量的算法和论文，并且能够解决人机博弈、自动驾驶等现实场景中的复杂问题。深度强化学习的代表作 AlphaGo，已经能够轻松战胜当今世界所有的人类围棋高手了。

相对于深度学习和强化学习，符号主义的 AI 应用表现虽然没有那么抢眼，但是仍有很大发展。近年来，大量知识图谱不断涌现，这些蕴含人类大量先验知识的宝库却尚未被深度学习有效利用。因此，在 AI 最新的发展趋势中，因果网络、知识图谱、深度学习、强化学习也呈现出相互融合的趋势。

在这一背景之下，来自 Google 的 DeepMind、谷歌大脑、MIT 等机构的 27 位作者在 2018 年发表了又一篇重磅论文《关系归纳偏置、深度学习和图网络》（Relational inductive biases, deep learning, and graph networks）。这篇论文指出人工智能的未来是把传统的贝叶斯因果网络和知识图谱，与深度强化学习融合。而把端到端学习与归纳推理相结合，将有望解决 Judea Pearl 所说的深度学习无法进行关系推理的问题，能进一步提升深度学习模型效果。

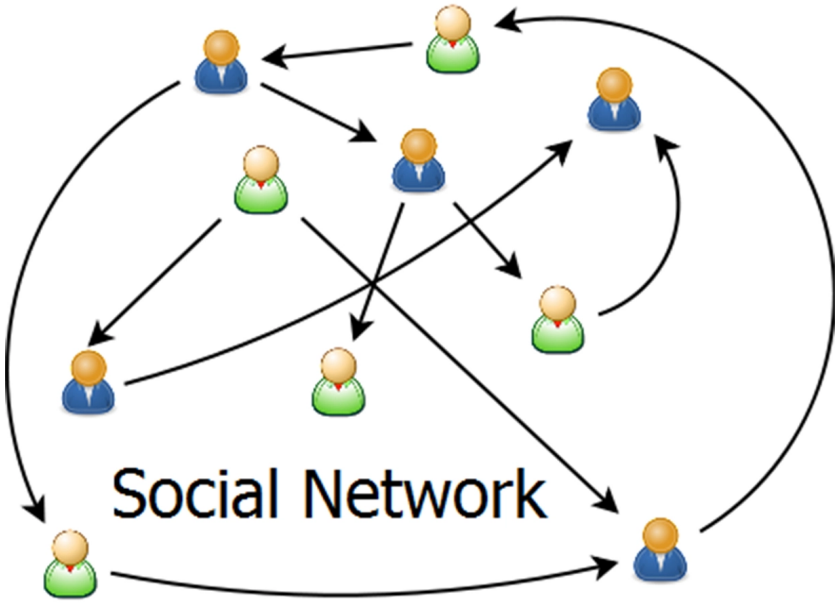
这样，以知识图谱为代表的符号主义、以深度学习为代表的连接主义，就逐渐脱离了原先各自独立发展的轨道，走上了协同并进的新道路。

具体来讲，怎么融合呢？DeepMind 认为要从图神经网络（Graph Neural Networks，GNN）入手。《关系归纳偏置、深度学习和图网络》这篇论文中也探讨了如何在深度学习结构（比如全连接层、卷积层和递归层）中，使用图神经网络（也叫图深度学习）的关系归纳偏置，为操纵结构化知识和生成结构化行为提供可能性。

那么，图神经网络和普通神经网络又有何不同？这就要从数据集的类型说起。

我们知道，神经网络已经在对象检测、图像识别、机器翻译和语音识别等领域大显身手，这是因为深度学习擅长捕捉欧几里得数据的隐藏模式，解决了计算机视觉和自然语言处理中的很多问题。像图像、文本、视频等尺寸维度确定并且有序排列的数据就是我们说的“欧几里得数据”。

对应于“欧几里得数据”，还有另一种类型的数据被称为“非欧几里德数据”，这类数据的对象关系非常复杂，并且具有相互依赖的图形，比如社交网络、知识图谱、复杂的文件系统等，它们的表现形式更加非结构化，难以用一个序列或者一个网络来表示，所以我们把非欧几里德数据也统称为图。下面就是一个复杂的社交网络图的抽象展示。



社交网络图

很明显，和简单的文本和图像相比，图更加复杂，我们处理它的难点有很多：

图的大小是任意的；

图的拓扑结构复杂，没有像图像一样的空间局部性；

图没有固定的节点顺序，没有固定的形式，而且无序节点可变大变小，其中节点可以有不同数量的邻居。换句话说，图没有一个参考节点，还常常是动态的；

图包含多模态的特征；

……

而现有的机器学习算法有一个核心假设，即实例彼此独立，这对于图数据来说是错误的，因为图数据的每个节点都通过各种类型的链接与其他节点相关。那么，对于这样的图数据，如何利用深度学习进行建模呢？

答案就是融合了因果推理、知识图谱、强化学习和深度学习等诸多元素的图神经网络。图神经网络的目标就是对图所描述的数据进行推理，建立起可以直接应用于图的神经网络，并提供方法来执行节点级、边级和图级预测任务，以完成卷积神经网络、循环神经网络等深度学习网络无法完成的任务。

因此，你不难看出，当神经网络发展到图神经网络，机器学习的应用范围就更可以进一步扩展了。在社交网络数据分析、推荐系统、物理建模、自然语言处理和图上的组合优化问题等，AI 应用肯定能有新的突破。

对于图神经网络，我们这里只是做拓宽视野性质的说明，如果你对它有兴趣，更多的内容可以参考清华大学整理的 GNN 论文集。相信在不久的将来，除了论文之外，还会出现更多好用、易用的 GNN 工具包供我们直接调用，解决更加具体的问题。

总结一下

好，我们这一讲到这里就结束了，我们现在总结一下。

在这一讲中，我从三个角度带你探讨了作为“机器学习的应用者”，如何利用好机器学习领域的最新进展，来赋能个人发展。

首先，通过迁移学习，我们可以利用各大公司开源、发布的最新、最稳定的模型，在其基础上进行微调，以适应自己的项目需要。

另外，我们还可以借助自动化机器学习，比如自动化特征工程、自动化模型选择和超参数调整、自动部署机器学习模型，来简化机器学习项目的开发部署流程，降低开发难度，提高我们的项目开发效率。

最后，作为机器学习的应用者，我们也应当了解人工智能从何处起源，要发展到哪里去，这将有助于我们洞悉未来，占得先机。而深度强化学习和图神经网络的出现，代表着符号主义、连接主义和行为主义这三大 AI 源流的汇聚，这种汇聚让最新的机器学习模型比以前更加强大，而且能应用到更多实际的场景中。

总之，机器学习这个领域异常庞大，发展也非常迅速，我们还有许许多多的热点问题没有探讨，比如 AI 和物联网与 5G 的结合、AI 和网络安全的结合，再比如对 AI 和伦理问题、AI 和数据安全、以及如何通过联邦学习在数据共享的同时保护数据隐私等等。如果你的工作和上述某些方面相关，或者对其中某个方向感兴趣，不妨去深入探索。

虽然我们的课程没有涉及每一个方向，但有一点是永恒不变的：**任何技术都需要“人”去实现它**。如果我们把 AI 看作是一个工具，那它只会让我们的工作变得更轻松、更愉悦。机器学习为我们提供了大量的机会，同时也提出了很多具体的问题和挑战。当然，问题和挑战本身都是机会，这取决于我们如何去应对它们。我希望你能多学习、多思考。

思考题

在这一讲的最后，我希望你能够分享一下你对机器学习的理解和感悟，以及你所了解的机器学习和整个人工智能领域的最新进展。最后，祝愿你在机器学习之路上走的更远，用好机器学习这个工具，并做出自己的探索和贡献！

极客时间

本讲重点

作为机器学习的“ 应用者 ”，如何跟上机器学习领域的最新进展？

首先，通过迁移学习，我们能够在自己的项目中利用上最新、最强的模型。第二，通过自动化机器学习能够让特征工程、模型选择和超参数调整，以及模型部署过程变得更为简单。第三，深度强化学习和图神经网络的出现意味着人工智能的应用领域变得越来越广阔，未来可期。

分享给需要的人，Ta订阅后你可得 20 元现金奖励

生成海报并分享

赞 2　提建议

© 版权归极客邦科技所有，未经许可不得传播售卖。 页面已增加防盗追踪，如有侵权极客邦将依法追究其法律责任。

上一篇　20 | 模型部署：怎么发布训练好的机器学习模型？

下一篇　一套习题，测出你对机器学习的掌握程度

精选留言 (1)

写留言

在路上
2021-10-15

佳哥好，短短21讲的内容，让我对机器学习有了初步而全面的认识。最大的收获是工具的强大，很多算法都有现成的类库，API也很简单，上手很容易。

展开

作者回复: 谢谢你一路上的分享和对课程的认可。“实用主义”——这正是我设计这个课程的出发点。对于工具的使用者而言，了解了工具的强大，是用好它的前提，对吧。💪♀

1